How well current saliency prediction models perform on UAVs videos?

Anne-Flore Perrin, Lu Zhang, Olivier Le Meur

How well current saliency prediction models perform on UAVs videos?*

Anne-Flore Perrin[1][0000−0003−4210−9517], Lu Zhang[2][0000−0002−8859−5453], and Olivier Le Meur[1][0000−0001−9883−0296]

[1] Univ Rennes, CNRS, IRISA, 263 Avenue Général Leclerc, 35000 Rennes, France
[2] Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France
anne-flore.perrin@irisa.fr, lu.ge@insa-rennes.fr, olemeur@irisa.fr

Abstract. It is exciting to witness the fast development of Unmanned Aerial Vehicle (UAV) imaging which opens the door to many new applications. In view of developing rich and efficient services, we wondered which strategy should be adopted to predict salience in UAV videos. To that end, we introduce here a benchmark of off-the-shelf state-of-the-art models for saliency prediction. This benchmark studies comprehensively two challenging aspects related to salience, namely the peculiar characteristics of UAV contents and the temporal dimension of videos. This paper enables to identify the strengths and weaknesses of current static, dynamic, supervised and unsupervised models for drone videos. Eventually, we highlight several strategies for the development of visual attention in UAV videos.

Keywords: Benchmark · Salience · Dynamic saliency models · Unmanned Aerial Vehicles (UAV) · Videos.

1 Introduction

Unmanned Aerial Vehicles (UAVs) are propitious for a broad range of applications. Drone racing tournaments, new autonomous delivery services, wildfire detection, and private house surveillance is a sample of these highly promising services. However, we reach the point that service improvement lies in technology-, imagery- and context-based solutions. Indeed, UAV imagery presents specificities, detailed later on, that are worth exploiting, especially when using videos.

Services mentioned above heavily rely on object detection, segmentation and even more on saliency detection. Salience expresses the extent of importance of an area, approximating perceptual processes of visual attention. It is represented as the prediction of the fixation probability density of a multimedia content [25]. Its prediction is essential in diverse fields, from content-aware re-targeting and compression to advertisement analyses. As UAV applications may further benefit

* The presented work is funded by the ongoing research project ANR ASTRID DISSOCIE (Automated Detection of SaliencieS from Operators' Point of View and Intelligent Compression of DronE videos) referenced as ANR-17-ASTR-0009.

from visual attention theory, we investigated whether state-of-the-art saliency models are suitable for predicting visual attention in UAV videos.

Several works already applied saliency prediction models in UAV services. An automatic salient object detection is implemented in [36] through mean-shift segmentation and edge-detection operators. In [37], the combination of frame alignment, forward/backward difference and blob detection creates a UAV video event summarization, tracking salient objects. Deep architectures are used for real-time autonomous indoor navigation [21, 11] and for monitoring wildfire [43], among others. Such approaches build upon basic deep-learning models, namely AlexNet [24] and CaffeNet [18].

Let us stress that only very few works address the temporal dimension in traditional and UAV videos. Methods that tackle the temporal dimension comprise hand-crafted motion features [10, 35], network architecture fed with optical flow [1], possibly in a two-layer fashion [1, 7], or Long Short-Term Memory (LSTM) architectures [2, 19, 40, 28] to benefit from their memory functionality.

Regarding conventional imaging, elegant, elaborate and efficient solutions have been developed for saliency prediction. However, we wonder if typical schemes keep their promises when dealing with UAV video characteristics. Indeed, this new-born imagery field is distinct from typical imagery in many aspects, including the bird-point-of-view which modifies the semantic and size of objects [23], the loss of pictorial depth cues [15] i.e. the lack of horizontal line [9], and the presence of camera movements [23].

In this paper, we provide answers to the above questions by first introducing the saliency models under review in Section 2. In Section 3, we go through the details of the benchmark. Results are discussed in Section 4. Finally, key takeaways of this benchmark are given in Section 5.

2 Saliency models

Tremendously different approaches have been developed throughout the years to reach high accuracy prediction of visual attention. We propose here a taxonomy to categorize salience solutions.

2.1 Taxonomy

With the introduction of temporal dimension in videos, Borji [3] proposed to classify models according to the use or not of motion features and deep learning architectures. We refine this taxonomy by considering all types of supervision and not only differentiating deep-learning-based models. Accordingly, we define five model categories:

- **Static Unsupervised (SU)**: Itti [17], LeMeur [27], GBVS [12], SUN [42], Judd [20], Hou [14], RARE2012 [34], BMS [41],
- **Static Deep learning (SD)**: Salicon [16], DeepNet [33], ML-Net [6], SalGAN [32],
- **Dynamic Unsupervised (DU)**: Fang [8], OBDL [13],

– **Dynamic Machine learning (DM)**: PQFT [10], Rudoy [35],
– **Dynamic Deep learning (DD)**: DeepVS [19], ACL-Net [40], STSconvNet [1], FGRNE [28].

Deep learning models could be further classified depending on their loss function(s) or architecture attributes (e.g. Convolutional Neural Networks (CNN) design, multi-level features, encoder-decoder system, LSTM(s)).

2.2 Models under study

Models under study were selected based on an exhaustive study of the state of the art, exhibiting most used models for cross-model comparisons. Accordingly, we evaluate a total of 11 off-the-shelf models, including 8 static models and 3 dynamic models. They are briefly described hereinafter, following our taxonomy categorization.

Static unsupervised models

– **Itti**: One of the first and most used static models is referred to as Itti's model. In [17], authors got inspired from primates' visual systems, especially the center-surround analytic behavior of retina and cortical lateral inhibition mechanisms. Accordingly, they investigated three modalities, namely color, intensity, and orientation, through features on variable scales.
– **Graph-Based Visual Saliency (GBVS)** [12]: GBVS is a bottom-up visual saliency model, exploiting the real-time ability of graph algorithms. It implements a feature extraction using biologically inspired filters, an activation measure by subtracting features at different scales and finally a normalization, applied based on local maxima, convolution, and non-linear weighting.
– **Saliency Using Natural statistics (SUN)** [42]: SUN mimics the visual system behavior to find potential targets by means of a Bayesian probabilistic framework. It includes computation of self-information, likelihood and location prior to reach an estimation of pointwise mutual information, which expresses the overall salience.
– **SIM** [30] : First, a convolution of the image with a bank of filters using a multi-resolution wavelet transform yields to a scale-space decomposition. Then, a Gaussian Mixture Model (GMM), trained on eye-fixation data, simulates the inhibition mechanisms in visual cortex cells. Finally, multi-scale information is integrated using an inverse wavelet transform.
– **RARE2012** [34]: The key of RARE2012 is its multi-scale rarity mechanism applied on low-level colour and medium-level orientation features extracted beforehand. To emphasise both local contrasts and global rare regions, highly salient regions have the lowest occurrence probabilities of pixel at all scales.
– **Hou** [14]: Assuming that sparsity in spatial and frequential domains discriminate the foreground from the background, the foreground is represented by the sign of the discrete cosine transform of the signal. The inverse transform of the three color channels foreground are squared, summed up and smoothed to get the saliency map.

– **Boolean Map Saliency (BMS)** [41]: First, a set of Boolean maps is created by applying uniformly distributed thresholds on the image color channels. Attention maps are derived from this set through morphological dilatation and Gaussian blurring. The final saliency map is the average attention map.

Dynamic machine learning model

– **PQFT** [10]: This model extends phase spectrum of Fourier transform models to a temporal approach. Chroma, luma and motion information form the quaternion of a frame. The quaternion sympletic form is converted to the frequency domain. Only the phase, which represents local information of the signal, is converted back to the spatial domain. Such information, after Gaussian filtering, produce the saliency map.

Static deep learning model

– **ML-Net** [6]: This multi-level Fully Convolutional neural Network (FCN) combines low- and high-level features to predict saliency. However, the true ingenuity of this model lies in the defined loss function, which penalises more errors occurring on salient pixels.

Dynamic deep learning models

– **DeepVS**: Jiang et al. [19] proposed an elegant architecture including two models applied sequentially. The Object-to-Motion CNN (OM-CNN) is encoding objectedness features through a complex and comprehensive network combining hierarchical spatial (coarse) and temporal saliency maps. To further compute dynamic salience of videos, a Saliency-Structured Convolutional LSTM (SS-ConvLSTM), is added. Its two successive LSTM networks leverage both short and long term correlations.
– **ACL-Net**: Wang et al. [40] implemented a CNN-LSTM architecture for video saliency. An attention module supervises the CNN, forcing it to learn static features and ensuring the LSTM to deal with dynamic characteristics.

3 Benchmark design

3.1 Baselines

To include comparison points for the above models, we include six baselines. First, the average saliency map over all observers and all sequences (OHM) brings out an overall representation of **Human Mean** (HM). Also, average saliency maps over observers for a sequence (SHM), and over all observers for all sequences but the one under study (abSHM) examine content-dependencies. Then, the **Center Bias** (CB) map is a centered isotropic Gaussian stretched to video frame aspect ratio [4]. Last, we add two **chance** representations, inspiring from [31]. We split SHMs into 16 blocks, which are redistributed in the final map following two 4x4 magic squares. Despite the lack of salience information, these shuffled maps cover a similar ratio and dynamic range than that of the ground truth.

3.2 Dataset

Specific datasets are required to conduct saliency performance analyses. Indeed, eye movements of human beings or any substitute, collected during stimuli visualization, establish the Ground Truth (GT) essential to assess the validity of predicted saliency maps.

EyeTrackUAV dataset [23] is the only UAV salience dataset available, to the best of our knowledge. It includes 19 different videos extracted from the UAV123 database [29]. These video sequences were captured from a fully stabilized and manually controlled off-the-shelf professional-grade UAV (DJI S1000) flying at low-altitudes (varying between 5-25 meters). Criteria for content selection were the diversity of environment, distance and angle to the scene, size of the principal object and presence of sky. Authors of [23] collected highly precise binocular gaze data (1000 Hz) from 14 subjects in free viewing conditions. Stimuli were videos with a resolution of 720p, 30 fps which represent overall 26599 frames and 887 seconds.

From data, two maps have been computed for evaluation purposes. Saliency maps were inferred directly from raw data gathered by the eye tracker. Binocular gaze recordings were averaged over all observers. These maps were then filtered using a Gaussian kernel and normalized [26]. Fixations were retrieved through a Dispersion-Threshold Identification (I-DT) algorithm [22]. Fixation points, in this spatiotemporal detection, are assumed to cluster together.

3.3 Metrics

To carry out the evaluation, we use six quality metrics included in the MIT benchmark [5, 26]: Correlation Coefficient (CC) ($\in [-1, 1]$), Similarity (SIM) ($\in [0, 1]$) the intersection between histograms of saliency, Area Under the Curve (AUC) Judd and Borji ($\in [0, 1]$), Normalized Scanpath Saliency (NSS) ($\in]-\infty, +\infty[$), and Kullback Leibler divergence (KL) ($\in [0, +\infty[$). Details of these metrics can be found in [26, 5]. Also, a very interesting comparison of metrics behavior is presented in [4].

4 Results and discussion

4.1 Analyses

Static, dynamic, non-supervised, machine-learning and deep-learning models are compared here. We discuss our results qualitatively, overall and on a frame-by-frame basis.

Qualitative analysis

The qualitative verification is done on the sequence *Person20*, which illustrates models efficiency in a typical scenario of UAV applications. By observing Figure 1, one can note that deep learning and dynamic models detected less salient areas than unsupervised static models. This is in line with the assumed sparsity of video saliency.

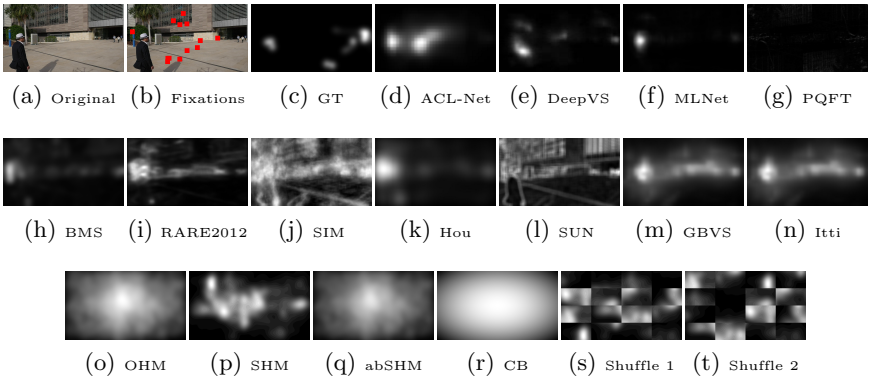


Fig. 1: Saliency maps of the 11 models, the ground truth, together with fixations and source of frame 1035 in *person20*.

Table 1: Performance of saliency models in average over all videos and all frames. Best performances are in bold. (AUC-B=AUC-Borji; AUC-J=AUC-Judd)

		CC ↑	SIM ↑	AUC-J ↑	AUC-B ↑	NSS ↑	KL ↓
DD	ACL-Net [40], 2018	0.4516	**0.3586**	0.8199	0.7717	1.9622	1.7803
	DeepVS [19], 2018	0.3986	0.3204	0.8059	0.7384	1.7904	1.9063
SD	MLNet [6], 2016	**0.4621**	0.3149	**0.8347**	**0.7866**	**2.1479**	**1.5857**
DM	PQFT [10], 2010	0.1367	0.1817	0.7054	0.5591	0.7364	2.3282
SU	BMS [41], 2016	0.3846	0.2482	0.8189	0.7855	1.8180	1.8053
	RARE2012 [34], 2013	0.3422	0.2566	0.7946	0.7582	1.5093	1.8240
	Hou [14], 2012	0.2811	0.2204	0.7565	0.7279	1.1529	1.9818
	SIM [30], 2011	0.2213	0.1893	0.7511	0.7235	0.9899	2.1482
	SUN [42], 2008	0.2065	0.1913	0.7361	0.7059	0.9593	2.1702
	GBVS [12], 2007	0.3687	0.2431	0.8125	0.7854	1.5915	1.8115
	Itti [17], 1998	0.3687	0.2431	0.8125	0.7855	1.5915	1.8115
Baselines	OHM	0.1889	0.1883	0.7208	0.6934	0.6881	2.3526
	SHM	0.1917	0.1944	0.7051	0.6790	0.6880	2.2463
	abSHM	0.1808	0.1868	0.7125	0.6840	0.6544	2.3539
	CB	0.2055	0.1980	0.7493	0.7172	0.7588	2.3776
	Shuffle 1	0.0282	0.1558	0.5724	0.5427	0.1315	2.4275
	Shuffle 2	-0.0085	0.1468	0.5323	0.5041	-0.0046	2.4745

Overall analysis

Overall results are presented in Table 1. The analysis relies on mean and standard deviation of raw results while Analysis of Variance (ANOVA) and multi-comparison studies have been conducted to verify the statistical significance of observed differences. We draw several observations.

Deep learning models present the best scores, especially for SIM for which differences with other models are statistically significant. This result was ex-

pected as the same behavior is observed since the development of deep learning techniques in saliency prediction [39, 16, 33, 6, 32].

Within static models, BMS model, followed closely by GBVS and Itti models, happens to be the most efficient static unsupervised model. SIM and SUN models are significantly less performing for CC, SIM, NSS and KL.

Regarding dynamic models, ACL-Net reaches the first and second position for all but AUC-Borji metric. DeepVS is also ranked high up behind the two other deep learning schemes. Contrarily, PQFT achieves statistically significantly the lowest performance of all models for all metrics and achieves similar performances as baselines. We thus recommend avoiding frequency-based solutions, even though this result must be replicated with other models and datasets for validation.

Baselines results show the potential of all models but PQFT - and SIM and SUN models with respect to KL and SIM - as they are proven statistically different from baselines. Despite being the most performing baseline, Center Bias results are quite low when compared to results on traditional contents. This raises suspicions about the suitability of this bias for UAV videos. We have indecisive results regarding overall image-dependent properties. Indeed, SHM is not statistically different from other HM baselines and is alternatively better (CC, SIM, KL) or worse (AUC-based and NSS).

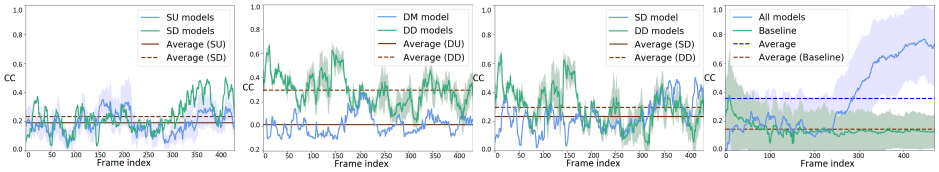
From all the above observations, UAV saliency models can be built upon static and dynamic deep learning models. Moreover, performances of today's solutions are fair and could potentially reach better scores through fine-tuning or training on a UAV-dedicated database.

Sequence-based Analysis

A challenge in video quality assessment is to design metrics that are representative of the entire sequence, considering the disparity in quality scores of video frames. Saliency prediction in videos presents the same challenge. In this benchmark, we tackled this issue by reporting average and according standard deviation values of metrics over the entire sequence. To dig further in UAV video saliency, let us discuss the accuracy of predictions on a frame-by-frame basis. This in-depth investigation exhibits potential perspectives for metrics development in video saliency prediction.

In Figures 2a to 2c, a comparison of averaged scores for models sorted by categories, following the above-defined taxonomy, is provided in terms of CC scores for sequence *bike3*. Also, overall CC results of all models and average baseline are reported in Figure 2d for *wakeboard10*.

High disparity. Overall, the disparity of standard deviation and mean values spreads out over a wide range among sequence frames. Hence, giving one score per video for saliency prediction may be fallacious. Also, we claim that the lack of consideration of temporal continuity in performance estimation may prevent thorough assessments of predictions. Implementing temporal coherence metric for video saliency prediction may be a solution to reach highly accurate and constructive analyses.



(a) SU vs SD *bike3* (b) DM vs DD *bike3* (c) SD vs DD *bike3* (d) All vs Baseline *wakeboard10*

Fig. 2: Temporal comparison of model categories in *bike3* and results of all models in *wakeboard10* for metric CC for all sequence frames. Colored envelop represents standard deviation associated with mean results of models.

Deep learning models perform better. Even on a frame-by-frame basis and amongst sequences presenting statistically different results, we can observe in Figures 2a and 2b the clear advantage of deep learning architectures over unsupervised and machine learning schemes, for both dynamic and static models.

Static vs dynamic deep learning models. There is no evidence of significant over-performance of static deep learning over dynamic deep learning approaches, as illustrated in Figures 2c. Actually, ANOVA has not rejected the null hypothesis when comparing the distributions of these two categories. We observe that results are content-dependent, hinting that the difference of performance between deep learning static and dynamic models lies in videos characteristics.

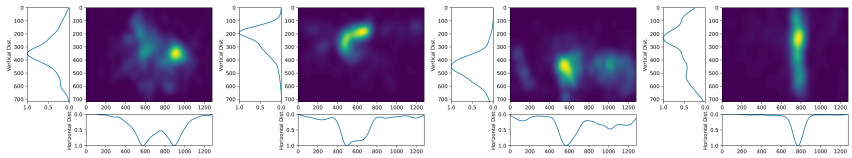
Event-related performance. We could not relate the varying features of EyeTrackUAV videos (overall angle and distance to the scene, environment, presence of sky, and object size) to models' efficiency. Our intuition is that model performances are event-related. We further studied our results to identify causes of abrupt changes in metrics along video frames.

It turns out that a decrease in performance may be caused by the entry of new objects of potential interest or environments variations. However, a point of interest arises in that a sudden increase in prediction accuracy often follows camera movements towards re-framing the content (change of camera angle, distance, speed or trajectory guided by an operator). Reframing reintroduces a center bias in the content. This explains the increase in prediction accuracy at the end of the scene *wakeboard10* in Figure 2d. This shows the importance of understanding patterns and biases in the used imagery.

4.2 Challenges and open problems

Several different aspects of our study must be examined and discussed, from the limitations due to the used dataset, to the "normalization" of metrics for better comparison and analyses of saliency prediction strategies.

Dataset. EyeTrackUAV presents a clear lack of non-natural videos. But most importantly, this dataset is not sufficient to train a deep model, even with data augmentation. To address the need of dedicated UAV video saliency prediction,



(a) bike3 (b) boat6 (c) building5 (d) car4

Fig. 3: Center bias for sequences *bike3*, *boat6*, *building5* and *car4*.

one needs new larger databases with more videos, possibly with natural and unnatural images.

Likewise, the larger the population taking part in the gaze data gathering is, the more accurate should be the ground truth - on which rely performance analyses. Also, based on our results, event-related annotations, even though complicated to collect, can provide a valuable contribution.

Finally, this dataset has been designed to study bottom-up attention. However, multiple UAV applications, such as aerial surveillance, monitoring and observation with drone, require the study of top-down attention, which is task-related. Both bottom-up and top-down attention ground truth is necessary for the development of UAV services.

Center bias. An additional UAV feature is that the center bias [38] is less significant in EyeTrackUAV sequences than in traditional video sequences, as seen in Figure 3. It may explain, in part, the weak performance of existing saliency models. Further analyses are required to confirm this compelling finding.

Metrics. As mentioned earlier, temporal metrics must be developed to correctly assess video saliency predictions. Our main point is that the averaged prediction performance over video frames may not be representative of the prediction along scenes in videos. It might be profitable to mimic the continuum of our visual gaze deployment. Several strategies may be deployed accordingly, possibly inspiring from event summarization [37].

With a slight shift of perspective, one may also consider to apply post-processing to predicted saliency maps to reach comparable metric results. To reduce metric biases, Kummerer et al. [25] proposed to turn a given fixation density into metric-specific saliency maps prior to metric computations.

5 Conclusion

UAV new services want to benefit from the substantial improvement in saliency prediction of this last decade. This benchmark reviews the ability of 11 state-of-the-art off-the-shelf prediction schemes to identify the direction to take toward the development of UAV imagery-dedicated models.

We studied qualitatively models performance over the EyeTrackUAV dataset which includes 19 natural UAV videos together with precise eye tracking information, collected from 14 subjects. Predicted maps have been assessed using

typical metrics, namely CC, SIM, AUC-Judd, AUC-Borji, NSS and KL. Overall scores of models are reported through mean and standard deviation over the entire sequence. ANOVA and multi-comparison analyses have been included to detect if models differences are statistically significant.

Several insights are provided among which are three key takeaway messages:

(1) In line with studies on traditional contents, static and dynamic deep learning models, trained on conventional contents, show the most promising results. This outcome is highly encouraging, especially with the potential of reaching higher performance through fine-tuning or training on larger UAV databases.

(2) Yet there are no significant difference between static and dynamic deep learning. This outcome has been shown to be content-dependent. Although video characteristics (angle, distance to the scene, environment type and object size) were not sufficient to explain our results, it seems that event-related annotations could help to efficiently learn saliency on such contents.

(3) The need of dedicated UAV-centric models is hereby made clear. We need to go deeper in content specificities and better encode them for future high-quality UAV-based services. Also, we found out that the center bias does not necessarily apply to UAV videos, which needs deeper exploration. A broad investigation of typical biases of attention and cognition could be carried out.

Different challenges have also been discussed, including the need to develop video-based metrics in view to further investigate video prediction performance as well as to better represent the quality of prediction along the entire sequence, or the necessity to create large datasets for UAV imagery saliency prediction.

References

1. Bak, C., Kocak, A., Erdem, E., Erdem, A.: Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Transactions on Multimedia **20**(7), 1688–1698 (2018)
2. Bazzani, L., Larochelle, H., Torresani, L.: Recurrent mixture density network for spatiotemporal visual attention. arXiv preprint arXiv:1603.08199 (2016)
3. Borji, A.: Saliency prediction in the deep learning era: An empirical investigation. arXiv preprint arXiv:1810.03716 (2018)
4. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(3), 740–757 (March 2019)
5. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: Mit saliency benchmark (2015)
6. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 3488–3493. IEEE (2016)
7. Dutt Jain, S., Xiong, B., Grauman, K.: Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3664–3673 (2017)
8. Fang, Y., Wang, Z., Lin, W., Fang, Z.: Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE Transactions on Image Processing **23**(9), 3910–3921 (Sep 2014)

9. Foulsham, T., Kingstone, A., Underwood, G.: Turning the world around: Patterns in saccade direction vary with picture orientation. Vision research **48**(17), 1777–1790 (2008)
10. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE transactions on image processing **19**(1), 185–198 (2010)
11. Guo, X., Cui, L., Park, B., Ding, W., Lockhart, M., Kim, I.: How will humans cut through automated vehicle platoons in mixed traffic environments? a simulation study of drivers' gaze behaviors based on the dynamic areas-of-interest (05 2018)
12. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in neural information processing systems. pp. 545–552 (2007)
13. Hossein Khatoonabadi, S., Vasconcelos, N., Bajic, I.V., Shan, Y.: How many bits does it take for a stimulus to be salient? In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
14. Hou, X., Harel, J., Koch, C.: Image signature: Highlighting sparse salient regions. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 194–201 (Jan 2012). https://doi.org/10.1109/TPAMI.2011.146
15. Howard, I.P., Rogers, B.: Depth perception. Stevens Handbook of Experimental Psychology **6**, 77–120 (2002)
16. Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 262–270 (2015)
17. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence **20**(11), 1254–1259 (1998)
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia. pp. 675–678. MM '14, ACM (2014)
19. Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: Deepvs: A deep learning based video saliency prediction approach. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 602–617 (2018)
20. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th international conference on computer vision. pp. 2106–2113. IEEE (2009)
21. Kim, D.K., Chen, T.: Deep neural network for real-time autonomous indoor navigation. arXiv preprint arXiv:1511.04668 (2015)
22. Krassanakis, V., Filippakopoulou, V., Nakos, B.: Eyemmv toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification. Journal of Eye Movement Research **7**(1) (Feb 2014)
23. Krassanakis, V., Perreira Da Silva, M., Ricordel, V.: Monitoring human visual behavior during the observation of unmanned aerial vehicles (uavs) videos. Drones **2**(4), 36 (2018)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
25. Kummerer, M., Wallis, T.S., Bethge, M.: Saliency benchmarking made easy: Separating models, maps and metrics. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 770–787 (2018)

26. Le Meur, O., Baccino, T.: Methods for comparing scanpaths and saliency maps: strengths and weaknesses. Behavior Research Method **45**(1), 251–266 (2013)
27. Le Meur, O., Le Callet, P., Barba, D.: Predicting visual fixations on video based on low-level visual features. Vision research **47**(19), 2483–2498 (2007)
28. Li, G., Xie, Y., Wei, T., Wang, K., Lin, L.: Flow guided recurrent neural encoder for video salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3243–3252 (2018)
29. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: European conference on computer vision. pp. 445–461. Springer (2016)
30. Murray, N., Vanrell, M., Otazu, X., Parraga, C.A.: Saliency estimation using a non-parametric low-level vision model. In: CVPR 2011. pp. 433–440 (June 2011)
31. Ninassi, A., Le Meur, O., Le Callet, P., Barba, D.: Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In: 2007 IEEE International Conference on Image Processing. vol. 2, pp. II–169. IEEE (2007)
32. Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.: Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081 (2017)
33. Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 598–606 (2016)
34. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., Dutoit, T.: Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. Signal Processing: Image Communication **28**(6), 642–658 (2013)
35. Rudoy, D., Goldman, D.B., Shechtman, E., Zelnik-Manor, L.: Learning video saliency from human gaze using candidate selection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1147–1154 (2013)
36. Sokalski, J., Breckon, T.P., Cowling, I.: Automatic salient object detection in uav imagery. Proc. of the 25th Int. Unmanned Air Vehicle Systems pp. 1–12 (2010)
37. Trinh, H., Li, J., Miyazawa, S., Moreno, J., Pankanti, S.: Efficient uav video event summarization. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 2226–2229. IEEE (2012)
38. Tseng, P.H., Carmi, R., Cameron, I.G., Munoz, D.P., Itti, L.: Quantifying center bias of observers in free viewing of dynamic natural scenes. Journal of vision **9**(7), 4–4 (2009)
39. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2798–2805 (2014)
40. Wang, Z., Ren, J., Zhang, D., Sun, M., Jiang, J.: A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing **287**, 68–83 (2018)
41. Zhang, J., Sclaroff, S.: Exploiting surroundedness for saliency detection: a boolean map approach. IEEE transactions on pattern analysis and machine intelligence **38**(5), 889–902 (2016)
42. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of vision **8**(7), 32–32 (2008)
43. Zhao, Y., Ma, J., Li, X., Zhang, J.: Saliency detection and deep learning-based wildfire identification in uav imagery. Sensors **18**(3), 712 (2018)